

Efficient Estimation of Quantiles in Missing Data Models

Iván Díaz ^{*1,2}

¹Division of Biostatistics and Epidemiology, Weill Cornell Medicine, New York, NY, USA.

²Google Inc., New York, NY, USA.

August 23, 2016

Abstract

We propose a novel targeted maximum likelihood estimator (TMLE) for quantiles in semiparametric missing data models. Our proposed estimator is locally efficient, \sqrt{n} -consistent, asymptotically normal, and doubly robust, under regularity conditions. We use Monte Carlo simulation to compare our proposed method to existing estimators. The TMLE is superior to all competitors, with relative efficiency up to three times smaller than the inverse probability weighted estimator (IPW), and up to two times smaller than the augmented IPW. This research is motivated by a causal inference research question with highly variable treatment assignment probabilities, and a heavy tailed, highly variable outcome. Estimation of causal effects on the mean is a hard problem in such scenarios because the information bound is generally small. In our application, the efficiency bound for estimating the effect on the mean is possibly infinite. This rules out \sqrt{n} -consistent inference and reduces the power for testing hypothesis of no treatment effect on the mean. In our simulations, using the effect on

*This research was completed during the author's tenure at Google Inc.; the manuscript was updated and revised after he joined Weill Cornell Medicine.

the median allows us to test a location-shift hypothesis with 30% more power. This allows us to make claims about the effectiveness of treatment that would have hard to make for the effect on the mean. We provide R code to implement the proposed estimators.

Key words: Quantile effects, information bound, \sqrt{n} -consistency, TMLE.

1 Introduction

Estimation of quantiles in missing data models is a statistical problem with applications to a variety of research areas, but has been somewhat overlooked in the semiparametric inference literature. For example, policy makers may be interested in evaluating the effect of an educational program on the tails of the skill distribution. In this case quantile treatment effects may be useful since they capture intervention effects that are heterogeneous across the outcome distribution. Quantiles may also be useful in economics research to compute inequality indicators such as the Gini coefficient. Quantile treatment effects may also be useful in the study of treatment effect heterogeneity, e.g., for assessing whether all quantiles are equally affected by treatment.

Our methods are motivated by an application to estimation of the causal effect of treatment on an outcome whose distribution exhibits heavy tails. The data we consider arises as part of various sales and services programs targeted to introduce new features to users of the AdWords advertisement platform at Google Inc. A important question for decision makers is to quantify the causal effect of these programs on the advertisers' spend through AdWords. The outcome we consider exhibits heavy tails, as there is a small but non-trivial number of advertisers who spend large quantities through AdWords. Heavy tailed distributions are often characterized by large or infinite variance, which in

turn yields a large or infinite efficiency bound for estimating the effect of treatment on the mean. As a consequence, the variance of all regular estimators is also large, possibly precluding \sqrt{n} -consistent inference and statistical significance at most plausible sample sizes. Therefore, though the effect on the mean is arguably an important parameter for this problem (the mean spend is directly related to total spend), \sqrt{n} -consistent inference for it may be impossible or very hard in our application. In Section 5 we present simulation results showing that the n -scaled mean squared error may not converge. This is a strong indication of an infinite efficiency bound.

Estimation of quantiles in missing data models may be of general interest in several scenarios, such as those described in the first paragraph of this introduction, irrespective of the skewness of the outcome. Our methods are motivated by using the effect on the quantiles as an alternative to the effect on the mean, as a test statistic for a location-shift hypothesis. We propose a novel targeted maximum likelihood estimator. This estimator is locally efficient in the non-parametric model and asymptotically linear, under certain regularity conditions. In our application, estimating a collection of quantiles of interest (e.g., 25%, 50% and 75%) allows us to make statements about treatment effects, even though we would have difficulty making similar statements for the mean, due to the large variability caused by the heavy tailed distribution.

Our goal is to estimate an unconditional quantile. An alternative goal is to estimate an outcome quantile conditional on the values of certain covariates. Though we do not estimate conditional quantiles, we use covariate information in order to correctly identify the unconditional quantiles under the missing at random assumption. Estimation of conditional quantiles is discussed, for example, by [Buchinsky \(1998\)](#); [Koenker \(2005\)](#); [Yu and Jones \(1998\)](#), among many others. Adaptation of our methods to estimate quantiles within strata of categorical baseline covariates are possible but are not discussed here.

In order to assess the performance of the proposed estimators in our real data application, we use Monte Carlo simulations based on a real dataset to approximate the bias, variance, mean squared error, and coverage probability of the confidence interval estimators for our application. Our proposed TMLE has the best performance across various modeling scenarios in comparison to the available alternative of IPW and augmented IPW (AIPW) estimation. We also use the simulation study to demonstrate that estimation of the effect on the median has improved power compared to the effect on the mean for testing the location-shift hypothesis.

Our paper is organized as follows. In Section 2 we discuss some existing approaches to estimation of quantiles in missing data models. In Section 3 we formally introduce the problem in terms of a closely related one: estimating the distribution function of an outcome missing at random. In Section 4 we present a summary of the available estimation methods, and present our proposed estimators for the quantiles of a variable missing at random as well as the effect of treatment on the quantiles, together with a theorem providing the conditions for asymptotic linearity and efficiency. In Section 5 we present two simulation studies, one with a synthetic data generating mechanism, and one using a real dataset from our motivating application. The Monte Carlo simulation study based on a real dataset is used to illustrate the performance of our estimator and show the benefits of using the median as a location parameter for the counterfactual distribution in the presence of heavy tails. Finally, in Section 6, we discuss some concluding remarks.

2 Related Work

Various methods exist that address the problem considered here. Wang and Qin (2010) consider pointwise estimation of the distribution function using the augmented inverse probability weighted estimator applied to an indicator function, where the missingness

probabilities and observed outcome distribution functions are estimated via kernel regression. They propose to use the distribution function to estimate the relevant quantiles using an outcome distribution estimator (i.e., the inverse of the estimated distribution function). Their approach suffers from various flaws stemming from the fact that the estimated distribution function may be ill-defined: direct inverse probability weighting may generate estimates outside $[0, 1]$, and pointwise estimation may yield a non-monotonic function. In addition, their approach may not be used in high dimensions since kernel estimators suffer from the curse of dimensionality.

[Zhao et al. \(2013\)](#) propose similar estimators for non-ignorable missing data, under the assumption that the missingness mechanism is linked to the outcome through a parametric model that can be estimated from external data sources. [Liu et al. \(2011\)](#), [Cheng and Chu \(1996\)](#), and [Hu et al. \(2011\)](#) consider estimators that yield estimated distribution functions in the parameter space, relying either on kernel estimators for the outcome distribution function, or knowledge of the true missingness probabilities. [Firpo \(2007\)](#) proposes to estimate the quantiles by minimizing an inverse probability weighted check loss function. Their estimator achieves non-parametric consistency by means of a propensity score estimated as a logistic power series whose degree increases with sample size. [Melly \(2006\)](#), [Frölich and Melly \(2013\)](#), and [Chernozhukov et al. \(2013\)](#) consider estimation of the quantiles under a linear parametric model for the distribution and quantile functions, respectively. Their methods lack the double robustness and efficiency properties of our proposal.

[Zhang et al. \(2012\)](#) propose a variety of methods, including an IPW and an AIPW. The AIPW estimator is expected to have similar asymptotic properties to our proposed TMLE, i.e., it is expected to be doubly robust, efficient, and asymptotically linear, under regularity conditions. In the context of estimation of the effect on the mean, targeted maximum likelihood estimators have consistently shown better performance than their

AIPW counterparts for finite samples (see e.g., [Porter et al., 2011](#)). In Section 5 we show, in simulation studies, that our proposed TMLE for the effect on the quantiles also has superior finite sample performance.

3 Notation and Estimation Problem

Let Y denote an outcome observed only when a missingness indicator M equals one, and let X denote a set of observed covariates satisfying $Y \perp\!\!\!\perp M \mid X$. We use P_0 to denote the true joint distribution of the observed data $Z = (X, M, MY)$. Assume we observe an i.i.d. sample Z_1, \dots, Z_n , and denote its empirical distribution by \mathbb{P}_n . We use the word *model* to refer to a set of probability distributions, and the expression *nonparametric model* to refer to the set of all distributions having a continuous density with respect to a dominating measure of interest. The word *estimator* is used to refer to a particular procedure or method for obtaining estimates of P_0 or functionals of it. We assume P_0 is in the nonparametric model \mathcal{M} , and use P to denote a general element of \mathcal{M} . For a function $h(z)$, we denote $Ph = \int h dP$. For simplicity in the presentation we assume that X is finitely supported but the results generalize to infinite support by replacing the counting measure by an appropriate measure whenever necessary. Under the assumption that $P_0(M = 1 \mid X = x) > 0$ almost everywhere, the distribution $F_0(y) \equiv Pr(Y \leq y)$ is identified in terms of P_0 as

$$\begin{aligned} F_0(y) &= \sum_x Pr_0(Y \leq y \mid X = x) Pr_0(X = x) \\ &= \sum_x Pr_0(Y \leq y \mid M = 1, X = x) Pr_0(X = x) \\ &= \sum_x G_0(y \mid 1, x) p_{X,0}(x), \end{aligned}$$

where we have denoted $G(y \mid 1, x) \equiv \Pr(Y \leq y \mid M = 1, X = x)$ and $p_X(x) \equiv \Pr(X = x)$. We use f to denote the density corresponding to F and $e(x)$ to denote $\Pr(M = 1 \mid X = x)$, following the convention in the propensity score literature. We also denote $\eta = (G, e)$. Consider the q -th quantile of the outcome distribution:

$$\theta = \inf\{y : F(y) \geq q\}.$$

We use the notation $\theta(P)$ to refer to the functional that maps an observed data distribution P into a real number. Given a consistent estimator \hat{G} of G_0 , the outcome distribution estimator $\hat{\theta}_{\text{OD}}$ is obtained as an (approximate) solution to the equation $\frac{1}{n} \sum_{i=1}^n \hat{G}(\theta \mid 1, X_i) = q$ is typically consistent, but it may not be \sqrt{n} -consistent. Various methods exist in the semi-parametric statistics literature that may be used to remedy this issue. The analysis of the asymptotic properties of such methods often relies on so-called von Mises expansions ([von Mises, 1947](#)) and on the theory of asymptotic lower bounds for estimation of regular parameters in semiparametric models (see, e.g., [Bickel et al., 1997](#); [Newey, 1990](#)).

The efficient influence function $D(Z)$ is one of the key concepts introduced by semi-parametric efficient estimation theory. This function characterizes all efficient, asymptotically linear estimators $\hat{\theta}$. Specifically, the following holds for any such estimator (see e.g., [Bickel et al., 1997](#)):

$$\sqrt{n}(\hat{\theta} - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n D(Z_i) + o_P(1/\sqrt{n}). \quad (1)$$

Asymptotic linearity allows the use of the central limit theorem to construct Wald-type asymptotically valid confidence intervals and hypothesis tests. For our target of inference θ , the efficient influence function in the non-parametric model is given below in [Lemma 1](#).

Lemma 1 (Efficient Influence Function). *The efficient influence function of θ at P in the*

non-parametric model is equal to

$$D(Z) = -\frac{1}{f(\theta)} \left[\frac{M}{e(X)} \{I_{(-\infty, \theta]}(Y) - G(\theta | 1, X)\} + G(\theta | 1, X) - q \right]. \quad (2)$$

When necessary, we favor the notation $D_{\eta, \theta}$ to indicate the dependence of D on the nuisance parameter $\eta = (G, e)$ and the quantile θ . Lemma 1 is a direct consequence of the functional delta method applied to the non-parametric estimator of $F_0(y)$, and the Hadamard derivative of the quantile functional given in Lemma 21.4 of [van der Vaart \(2000\)](#). Note that if there is no missingness, then $P(M = 1) = 1$, and $D(Z)$ reduces to $D(Z) = -(I_{(-\infty, \theta]}(Y) - q)/f(\theta)$. Then (1) is the standard asymptotic linearity result for the sample median (see, e.g., corollary 21.5 of [van der Vaart, 2000](#)).

Lemma 2 (Double Robustness of $D_{\eta, \theta}$). *Let $\eta = (G, e)$ with either $G = G_0$ or $e = e_0$. Then $P_0 D_{\eta, \theta_0} = 0$.*

In the above lemma we established the double robustness of the efficient influence function. As a consequence of this lemma, under standard conditions for the analysis of M -estimators (e.g., Theorem 5.9 of [van der Vaart, 2000](#)), an estimator $\hat{\theta}$ that satisfies $\mathbb{P}_n D_{\hat{\eta}, \hat{\theta}} = 0$ is consistent if either \hat{G} or \hat{e} is consistent, but not necessarily both. This argument motivates the construction of an AIPW estimator as the (approximate) solution to $\mathbb{P}_n D_{\hat{\eta}, \theta} = 0$ in θ , for auxiliary estimators \hat{G} and \hat{e} . This estimator was originally proposed by [Zhang et al. \(2012\)](#). Specifically, the estimator is defined as an approximate solution to the equation

$$\frac{1}{n} \sum_{i=1}^n \left[\frac{M_i}{\hat{e}(X_i)} \{I_{(-\infty, \theta]}(Y_i) - \hat{G}(\theta | 1, X_i)\} + \hat{G}(\theta | 1, X_i) \right] = q \quad (3)$$

in θ . We denote this estimator with $\hat{\theta}_{\text{AIPW}}$. Similarly, [Zhang et al. \(2012\)](#) proposed to

estimate θ_0 as the value $\hat{\theta}_{\text{IPW}}$ that solves the equation

$$\sum_{i=1}^n \frac{M_i}{\hat{e}(X_i)} I_{(-\infty, \theta]}(Y_i) = q$$

in θ . In our simulation study of Section 5 we also consider the inverse probability weighted estimator proposed by [Firpo \(2007\)](#), defined as

$$\hat{\theta}_{\text{FIRPO}} = \arg \max_{\theta} \sum_{i=1}^n \frac{M_i}{\hat{e}(X_i)} (Y_i - \theta)(I_{(-\infty, \theta]}(Y_i) - q),$$

for an estimate \hat{e} of e_0 .

4 Targeted Maximum Likelihood Estimator

As an alternative to the above estimators, in this section we propose a method to construct an estimator \tilde{P} , and estimate θ_0 with the substitution estimator $\hat{\theta}_{\text{TMLE}} = \theta(\tilde{P})$. Our proposal is such that the component $\tilde{\eta}$ of \tilde{P} satisfies

$$\mathbb{P}_n D_{\tilde{\eta}, \hat{\theta}_{\text{TMLE}}} = o_P(1/\sqrt{n}). \quad (4)$$

Using M -estimation and empirical process theory we derive the conditions under which this estimator is consistent, efficient, and asymptotically normal. We present the proposed estimation algorithm along with theoretical results establishing its asymptotic properties. In our simulation studies of Section 5, we use synthetic and real datasets to illustrate the superior finite-sample performance of our estimator in comparison to the above competitors.

Targeted maximum likelihood estimation is a general estimation method concerned with the construction of substitution estimators that solve a target estimating equation.

The construction of a TMLE is carried out in three steps as follows. First, estimate the nuisance parameter η by means of standard (possibly data-adaptive) prediction techniques. Second, propose a parametric fluctuation submodel to iteratively tilt the initial estimators towards a solution of the target estimating equation. This submodel is such that its score spans the components of the estimating function; its parameter is estimated through standard maximum likelihood techniques. Because maximum likelihood estimates solve the score equation, it follows that the desired estimating equation is solved. Since the TMLE increases the likelihood of the initial estimators in the direction of the efficient influence function, it has been conjectured that it has improved finite sample properties, compared to the AIPW estimator. Such improvements have been illustrated for causal effects on the mean in simulation studies by [Gruber and van der Laan \(2010\)](#); [Porter et al. \(2011\)](#); [Stitelman et al. \(2012\)](#), among others, and are illustrated in this article for causal effects on the quantiles.

When the efficient influence function estimating equation is used to construct the TMLE, the resulting estimator enjoys the same asymptotic properties as the standard AIPW (e.g., double robustness, efficiency), under standard regularity conditions.

The reader interested in further discussion and other technical details underlying the general TMLE methodology is referred to [van der Laan and Rubin \(2006\)](#) and [van der Laan and Rose \(2011\)](#). We now proceed to define the estimator for our problem. Consider the following iterative procedure:

1. *Initialize.* Obtain initial estimates \hat{e} and \hat{G} of e_0 and G_0 . We discuss possible options to estimate these quantities in [Section 4.1](#) below.

2. *Compute $\hat{\theta}$.* For the current estimate \hat{G} , compute

$$\hat{F}(y) = \frac{1}{n} \sum_{i=1}^n \hat{G}(y \mid 1, X_i),$$

and $\hat{\theta} = \inf\{y : \hat{F}(y) \geq q\}$.

3. *Update \hat{G} .* Let \hat{g} denote the density associated to \hat{G} , and consider the exponential submodel

$$\hat{g}_\epsilon(y \mid 1, x) = c(\epsilon, \hat{g}) \exp\{\epsilon H_{\hat{\eta}, \hat{\theta}}(z)\} \hat{g}(y \mid 1, x),$$

where $c(\epsilon, \hat{g})$ is a normalizing constant and

$$H_{\hat{\eta}, \hat{\theta}}(z) = \frac{1}{\hat{e}(x)} \{I_{(-\infty, \hat{\theta}]}(y) - \hat{G}(\hat{\theta} \mid 1, x)\}$$

is the score of the model. Estimate ϵ as

$$\hat{\epsilon} = \arg \max_{\epsilon} \sum_{i=1}^n M_i \log \hat{g}_\epsilon(Y_i \mid 1, X_i).$$

The updated estimator of g_0 is given by $\hat{g}_{\hat{\epsilon}}$.

4. *Iterate.* Let $\hat{g} = \hat{g}_{\hat{\epsilon}}$ and iterate steps 2-3 until convergence.

The TMLE of θ_0 is denoted by $\hat{\theta}_{\text{TMLE}}$ and is defined as $\hat{\theta}$ in the last iteration. We also use \tilde{P} to denote the estimate of P_0 obtained in the last iteration.

The optimization problem in step 3 above is convex in ϵ , so that under regularity conditions we expect the algorithm to converge to the global optimum. In our simulation studies we found it practical to stop once $|\hat{\epsilon}| < 10^{-4} \times n^{-0.6}$. Note that the MLE of ϵ in

the TMLE algorithm satisfies

$$\sum_{i=1}^n \frac{M_i}{\hat{e}(X_i)} \{I_{(-\infty, \hat{\theta}_{\text{TMLE}}]}(Y_i) - \tilde{G}(\hat{\theta}_{\text{TMLE}} \mid 1, X_i)\} = o_P(1/\sqrt{n}). \quad (5)$$

This, together with the stopping criteria and the definition of the TMLE as

$$\frac{1}{n} \sum_{i=1}^n \tilde{G}(\hat{\theta}_{\text{TMLE}} \mid 1, X_i) = q,$$

yields $\mathbb{P}_n D_{\tilde{\eta}, \hat{\theta}_{\text{TMLE}}} = o_P(1/\sqrt{n})$. The following theorem, proved in Appendix A, establishes the consistency, asymptotic normality, and efficiency of the TMLE.

Theorem 1 (Asymptotic Distribution of $\hat{\theta}_{\text{TMLE}}$). *Let $\hat{\theta}_{\text{TMLE}}$ and \tilde{G} denote the TMLE of θ_0 and G_0 as defined above. Let $\|f\|^2 = \int f dP_0$ denote the squared $L_2(P_0)$ norm. Denote $h_{0,\theta}(x) = G_0(\theta \mid 1, x)$ and $\tilde{h}_\theta(x) = \tilde{G}(\theta \mid 1, x)$, and assume*

(i) $\tilde{\eta} = (\tilde{G}, \hat{e})$ converges to some $\eta_1 = (G_1, e_1)$ in the sense that $\|\tilde{h}_{\theta_0} - h_{1,\theta_0}\| \|\hat{e} - e_1\| = o_P(1/\sqrt{n})$, with either $h_{1,\theta_0} = h_{0,\theta_0}$ or $e_1 = e_0$.

(ii) $P_0 D_{\tilde{\eta}, \theta_0}$ is an asymptotically linear estimator of the map $\eta \rightarrow P_0 D_{\eta_1, \theta_0}$ at $\eta = \eta_1$.

(iii) The class of functions $\{D_{\eta, \theta} : |\theta - \theta_0| < \delta, \|h_\theta - h_{1,\theta}\| < \delta, \|e - e_1\| < \delta\}$ is Donsker for some $\delta > 0$ and such that $P_0(D_{\eta, \theta} - D_{\eta_1, \theta_0})^2 \rightarrow 0$ as $(\eta, \theta) \rightarrow (\eta_1, \theta_0)$.

Then $\sqrt{n}(\hat{\theta}_{\text{TMLE}} - \theta_0) \rightarrow N(0, \sigma^2)$. If $(\tilde{G}, \hat{e}) = (G_0, e_0)$, then $\hat{\theta}_{\text{TMLE}}$ is efficient with $\sigma^2 = \text{Var}\{D_{\eta_0, \theta_0}(Z)\}$.

4.1 Initial Estimators

Assumption (i) of Theorem 1 requires that at least one of G_0 or e_0 is consistently estimated at a fast enough rate. When the number of covariates is large, the curse of dimensionality

precludes the use of non-parametric estimators for these parameters. A common approach is to make use of parametric working models to estimate the nuisance parameters. Unfortunately, these parametric models often fail to describe the complex relations arising in large dimensional data sets and may therefore invalidate the conclusions of an otherwise well designed study (Starmans, 2011). In these scenarios, we advocate for the use flexible, data adaptive estimators to fit these quantities, so that assumption (i) remains plausible. One such approach is given by Super Learning (van der Laan et al., 2007). Super Learning is an ensemble learning algorithm that works in three steps as follows. First, a library of candidate estimators is proposed. This library usually contains many flexible estimation algorithms, and may contain some less flexible algorithms often hypothesized by subject-matter experts based on a-priori scientific knowledge. Second, the data is randomly split in a number of validation and training sets. Each algorithm is then trained in each training set, with its predictive performance estimated using the validation set. Lastly, the estimated predictive performance of the prediction algorithms is used to estimate their weights in a weighted convex combination of predictive algorithms. The super learner is defined as the resulting convex combination of candidate algorithms.

The super learner algorithm, discussed by Polley et al. (2011) in the context of regression, may be used to estimate the probabilities e_0 . Estimation of a conditional expectation is a problem extensively addressed in the statistical learning literature and we omit further discussion here. In contrast, data-adaptive estimation of a conditional density is a problem that has enjoyed considerably less attention. In Appendix B we discuss a super learning method to estimate the density g_0 of Y conditional on $(M = 1, X = x)$.

4.2 Estimating the Causal Effect on the Treated

In this subsection we discuss estimation of the causal effect of treatment on an outcome quantile among the treated. Specifically, let X denote a set of pre-treatment variables, let T denote a binary variable indicating the treatment group, and let Y denote the outcome of interest. We define the potential outcomes $Y_t : t \in \{0, 1\}$ as the outcome that would have been observed if, contrary to the fact, $P(T = t) = 1$. We assume that (i) $T \perp\!\!\!\perp Y_0 \mid X$, and that (ii) $e(x) = P(T = 1 \mid X = x) < 1$ almost everywhere. Assumption (i) is often referred to as the *no unmeasured confounders* or *ignorability* assumption, and states that all factors that are simultaneous causes of T and Y must be measured. Assumption (ii) is referred to as the *positivity* assumption, and ensures that all units have a non-zero chance of falling in the control arm $T = 0$ so that there is enough experimentation. Note that $Y_1 = Y$ on the event $T = 1$, so that the q -th quantile of Y_1 among units with $T = 1$ may be optimally estimated by the sample quantile of Y among treated units. Thus, we focus our attention on estimation of the quantile of Y_0 among units with $T = 1$. Let $F(y) = P(Y_0 \leq y \mid T = 1)$ denote the distribution function of Y_0 conditional on $T = 1$, then our target estimand is given by

$$\theta = \inf\{y : F(y) \geq q\}.$$

Under assumptions (i) and (ii) above, the distribution function F identified as

$$F(y) = \sum_x P_Y(y \mid 0, x) p_X(x \mid 1),$$

where $G(y \mid 0, x) = Pr(Y \leq y \mid T = 0, X = x)$ and $p_X(x \mid 1) = Pr(X = x \mid T = 1)$. The efficient influence function for estimation of θ in the non-parametric model may be found

using similar techniques as in the previous section as

$$D(Z) = -\frac{1}{f(\theta)} \left[\frac{1-T}{E(T)} \frac{e(X)}{1-e(X)} \{I_{(-\infty, \theta]}(Y) - G(\theta | 0, X)\} + \frac{T}{E(T)} \{G(\theta | 0, X) - q\} \right], \quad (6)$$

where f is the probability density function associated to F .

The targeted maximum likelihood estimation algorithm involves the following steps:

1. *Initialize.* Obtain initial estimates \hat{e} and \hat{G} of e_0 and G_0 .
2. *Compute $\hat{\theta}$.* For the current estimate \hat{G} , compute

$$\hat{F}(y) = \frac{1}{\sum_i T_i} \sum_{i=1}^n T_i \hat{G}(y | 0, X_i),$$

and $\hat{\theta} = \inf\{y : \hat{F}(y) \geq q\}$.

3. *Update \hat{G} .* Let \hat{g} denote the density associated to \hat{G} , and consider the exponential model $\hat{g}_\epsilon(y | 0, x) = c(\epsilon, \hat{g}) \exp\{\epsilon H_{\hat{\eta}, \hat{\theta}}(z)\} \hat{g}(y | 0, x)$, where $c(\epsilon, \hat{g})$ is a normalizing constant and

$$\hat{H}_{\hat{\eta}, \hat{\theta}}(z) = \frac{\hat{e}(X)}{1 - \hat{e}(x)} \{I_{(-\infty, \hat{\theta}]}(y) - \hat{G}(\hat{\theta} | 0, x)\}$$

is the score of the model. Estimate ϵ as $\hat{\epsilon} = \arg \max_{\epsilon} \sum_{i=1}^n (1 - T_i) \log \hat{g}_\epsilon(Y_i | 0, X_i)$.

The updated estimator of g is given by $\hat{g}_{\hat{\epsilon}}(y | 0, x)$.

4. *Iterate.* Let $\hat{g} = \hat{g}_{\hat{\epsilon}}$ and iterate steps 2-3 until convergence.

The TMLE of θ_0 is denoted by $\hat{\theta}_{\text{TMLE}}$ and is defined as the value of $\hat{\theta}$ in the last iteration.

Arguing as in the proof for Theorem 1 we find that this TMLE is asymptotically linear, doubly robust, and locally efficient, under regularity conditions.

5 Simulation and Case Studies

5.1 Synthetic Data Simulation

In an controversial paper, [Kang and Schafer \(2007\)](#) conducted a set of simulations to study the performance of doubly robust estimators for a missing data problem under misspecification of the outcome and treatment models. In particular, they focus on a situation where the weights $M_i/\hat{e}(X_i)$ are highly variable; a situation in which standard AIPW estimators generally have poor performance (see also [Robins et al., 2007b](#)). This simulation setup was further considered by [Porter et al. \(2011\)](#) with the objective of assessing the performance of various targeted maximum likelihood estimators.

In this section we revisit the simulation of [Kang and Schafer \(2007\)](#) with the modified objective of estimating the causal effect of a treatment variable on the median of the potential outcomes. The data is generated as follows. Let W_1, \dots, W_4 be independent normally distributed variables with mean zero and variance one. The treatment variable T is generated from a Bernoulli distribution with probability equal to $\text{expit}(-W_1 + 0.5 \times W_2 - 0.25 \times W_3 - 0.1 \times W_4)$, where $\text{expit}(x) = \{1 + \exp(-x)\}^{-1}$. The outcome is generated as $Y = 210 + 27.4 \times W_1 + 13.7 \times W_2 + 13.7 \times W_3 + 13.7 \times W_4 + N(0, 1)$. From this, we can determine the effect of T on the median of Y as $\text{median}(Y_1) - \text{median}(Y_0) = 0$. In average, 50% of the units are treated, with treatment probabilities in the interval $[0.01, 0.98]$. Estimators of the propensity score e_0 are therefore expected to lead to inverse probability weights with high variability, a situation in which doubly robust estimators may perform poorly. The

researcher does not observe the covariates W , but only the following transformations:

$$X_1 = \exp(W_1/2)$$

$$X_2 = W_2/\{1 + \exp(W_1)\} + 10$$

$$X_3 = (W_1 \times W_3/25 + 0.6)^3$$

$$X_4 = (W_2 + W_4 + 20)^2.$$

As it is bound to happen under standard practice using standard parametric models, inconsistent estimation of the outcome and treatment mechanisms occurs when the observed variables X are used to fit linear (logistic) regression estimators. Consistent estimation is achieved when the transformations $W(X)$ implied by the previous display are used instead. We consider four modeling scenarios: (a) \hat{G} and \hat{e} are consistent; (b) \hat{G} is consistent but \hat{e} is not; (c) \hat{e} is consistent but \hat{G} is not; and (d) both \hat{G} and \hat{e} are inconsistent. We generate 1000 datasets for each sample size $n \in \{200, 500\}$, and compute the TMLE, AIPW, IPW, Firpo, and OD estimators for each dataset. We then use the average and standard deviation of the 1000 estimates to approximate the bias and MSE of the estimators. In the simulation, \hat{G} is a normal distribution with the corresponding estimated conditional mean and residual variance. The code used to carry out this simulation is presented in [Appendix C](#).

Results The results of the simulation are presented in [Table 1](#). The TMLE is seen to have better performance in terms of MSE than all competitors across scenarios. This illustrates our conjectured improved finite sample performance. In scenario (a), the TMLE and AIPW are expected to have the same asymptotic distribution. This is corroborated in this simulation at sample size $n = 500$. In all other scenarios, the TMLE and AIPW may have different asymptotic distributions. When the treatment mechanism is misspecified,

Scenario	Estimator	$n = 100$			$n = 500$		
		$\sqrt{\text{MSE}}$	Bias	SD	$\sqrt{\text{MSE}}$	Bias	SD
(a)	TMLE	1.33	-0.00	1.33	0.71	0.01	0.71
	AIPW	1.42	0.01	1.42	0.71	-0.00	0.71
	IPW	9.59	-0.13	9.59	3.58	0.02	3.58
	Firpo	7.65	-0.11	7.65	3.02	0.12	3.01
	OD	0.29	0.02	0.29	0.11	-0.01	0.11
(b)	TMLE	1.33	0.01	1.33	0.70	0.00	0.70
	AIPW	1.42	-0.00	1.42	0.70	-0.00	0.70
	IPW	10.65	-5.32	9.23	6.34	-5.38	3.36
	Firpo	12.56	-2.64	12.28	14.92	0.96	14.89
	OD	0.29	0.02	0.29	0.11	-0.01	0.11
(c)	TMLE	6.13	-1.50	5.94	2.63	-0.33	2.61
	AIPW	7.18	-0.95	7.12	2.98	-0.22	2.98
	IPW	9.59	-0.13	9.59	3.58	0.02	3.58
	Firpo	7.65	-0.11	7.65	3.02	0.12	3.01
	OD	8.33	-7.21	4.16	7.68	-7.46	1.82
(d)	TMLE	8.40	-5.06	6.71	5.37	-4.44	3.01
	AIPW	8.92	-5.04	7.36	5.54	-4.72	2.89
	IPW	10.65	-5.32	9.23	6.34	-5.38	3.36
	Firpo	12.56	-2.64	12.28	14.92	0.96	14.89
	OD	8.33	-7.21	4.16	7.68	-7.46	1.82

Table 1: Kang and Schafer simulation results. MSE is the mean squared error and SD is the standard deviation.

the bias of the IPW and Firpo's estimators does not disappear as n increases. In addition, these estimators have a much larger variability than the doubly robust estimators TMLE and AIPW. The OD estimator has a much better performance than all other estimators in scenarios (a)-(b), when the outcome distribution is correctly estimated. This is a well known fact stemming from the fact that the variance of the MLE in a parametric model is smaller than the non-parametric efficiency bound. The OD estimator lacks the double robustness property and is therefore inconsistent in scenarios (c)-(d).

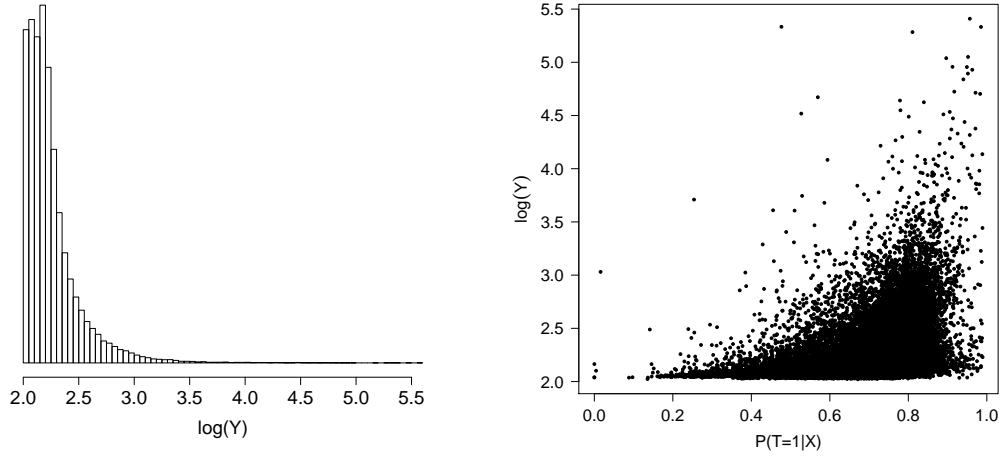
5.2 Real Data Simulation

In this section we illustrate the finite sample performance of the estimators in our motivating example. We estimate the effect on the quantiles among the treated, which is defined in the previous section. We use data from one of the AdWords programs at Google to create a data generating mechanism that mimics key features of our motivating applications such as high-dimensionality and heavy-tailed outcomes. We assess the performance of our estimators using these data generating mechanisms, which gives us a better idea of the expected performance in our real datasets, in comparison with synthetically generated data.

In our motivating problem, treatment consists of proactive consultations by sales representatives that help identify advertisers' business goals and suggest changes to improve performance. Since advertisers do not always adopt the proposed changes, a unit is considered treated if it is offered and accepts treatment. As a result, treatment is not randomized and we must use methods for observational data to assess the effect of these programs. The original dataset consists of 40,303 units, with 29,362 being treated. To adjust for confounders of the relation between treatment and spend through AdWords, we use 93 variables containing baseline characteristics of the customer as well as activity on their AdWords account.

The sample size used in the simulations is $n = 5,000$. This is a relatively small sample size compared to the number of covariates. Though admittedly smaller than the typical sizes seen in our applications, a sample size of 5,000 serves the present purpose of illustrating the finite sample performance of the estimators, while allowing for a computationally feasible simulation.

We have standardized the outcome to a variable with mean 10 and standard deviation 5 before carrying out our analyses. These values are selected arbitrarily and do not reflect



(a) Histogram of the natural logarithm of the standardized outcome. (b) Log-outcome as a function of the propensity score.

any particular feature of the data. Figure 1a shows the distribution of the logarithm of the standardized outcome, which can be seen to exhibit heavy tails and a large variability, even in the logarithmic scale. In addition, Figure 1b shows that larger values of the outcome are associated with larger propensities to receive treatment. This is because, in our application, larger customers are more likely to receive a sales consultation by a representative. Though it makes practical sense from a business perspective, this poses an additional challenge for estimation of causal effects since larger values of the outcome are associated with smaller control probabilities. This greatly increases the non-parametric efficiency bound for estimation of the effect on the mean (see [Robins et al., 2007a](#), for further discussion on this issue).

We use a re-sampling scheme based on parametric fits to the data in order to recreate a scenario that closely resembles the real data generating mechanism, while still allowing us to assess the performance of the estimators under different types of model misspecification with feasible computation times. We simulated 1,000 datasets from the observed data as

follows. First, we fit a logistic regression with main terms to the probability of treatment conditional on the covariates on the real datasets. We then treat this logistic regression function as the true probability of treatment conditional on covariates. Second, we fit a main terms quantile regression to the outcome, separately for the control and the treated group, for 500 quantiles, using the `quantreg` R package (Koenker, 2013). The resulting quantile function is then treated as the true outcome distribution conditional on treatment and covariates. We generate a sample by first drawing covariates from the empirical distribution (i.e., sampling with replacement). We then use the above probability of treatment conditional on covariates to draw a treatment indicator, and the above distribution of the outcome conditional on treatment and covariates to draw an outcome value. We are interested in estimating the effect of treatment on the 25%, 50%, and 75% quantiles of the outcome distribution. The true values for our data generating mechanism are 0.10, 0.13, and 0.23, respectively.

We compare the performance of the estimators in the same four scenarios resulting from the correct or incorrect misspecification of the estimators for e_0 and G_0 considered in Section 5.1. Misspecification of the estimators is carried out by omitting 30 of the observed covariates when fitting the initial estimators. The omitted covariates are chosen at random but fixed through the simulation. Estimation of \hat{G} is carried out by fitting a parametric quantile regression algorithm on 500 equally spaced quantiles using the R package `quantreg`. As a result, the initial density estimate \hat{g} has point mass $1/500$ at each of the initial quantiles, and the effect of the MLE in step 3 of Section 4 is to update the probability mass of each point. This algorithm is implemented in the R code presented in Appendix C.

Estimator performance is assessed in terms of percent bias, variance, mean squared error (MSE). For each generated dataset, we estimated the effect of treatment on the 25%,

50%, and 75% quantiles using each of the estimators. We then approximate the bias, variance, and MSE using empirical means across the 1,000 simulated datasets. The results are presented in Table 2.

Results The TMLE outperforms all its competitors in terms of MSE, with the exception of the OD estimator when the outcome distribution is consistently estimated. Though the OD estimator is more efficient than the TMLE if the outcome distribution is estimated consistently, it lacks the double robustness property. In general, the OD estimator will only be \sqrt{n} -consistent in the unlikely case that the outcome distribution is consistently estimated in a parametric model.

We conjecture that the improved performance of the TMLE over the AIPW is due to the property of the TMLE that it increases the likelihood of the estimate \hat{G} in the direction of the efficient influence function, and therefore provides a better bias-variance trade-off. In addition to having smaller variance, the TMLE has consistently a smaller bias across simulation scenarios.

There are important efficiency gains obtained by using the TMLE in comparison to its competitors. For example, for $q = 0.5$ and scenario (i), the TMLE can deliver the same precision as $\hat{\theta}_{\text{FIRPO}}$ using 63% fewer sample units. Similarly, the TMLE provides important efficiency gains compared to the AIPW. For example, in scenario (iii) and $q = 0.75$, the TMLE attains a performance comparable to the AIPW with 83% fewer sample units.

5.3 Testing The Location-shift Hypothesis

In our motivating example, consider the location-shift hypothesis that $p_{Y_1}(y \mid T = 1) = p_{Y_0}(y - \beta \mid T = 1)$, where p_{Y_t} denotes the density function of the counterfactual variable Y_t . If $\beta > 0$, this hypothesis tells us that treatment had a positive effect on the outcome by shifting the distribution of spend through AdWords to the right. We are interested in

Scen.	Estim.	$q = 0.25$			$q = 0.5$			$q = 0.75$		
		RMSE	Bias(%)	SD	RMSE	Bias(%)	SD	RMSE	Bias(%)	SD
(a)	TMLE	1.00	4.73	0.026	1.00	0.62	0.029	1.00	-2.86	0.077
	AIPW	1.16	8.96	0.030	2.14	6.38	0.062	1.01	0.73	0.078
	IPW	1.95	-2.82	0.052	3.16	-5.71	0.092	1.60	-17.35	0.117
	Firpo	1.38	-0.48	0.037	1.63	-1.14	0.048	1.42	0.14	0.110
	OD	0.83	-8.47	0.021	1.03	-9.71	0.028	0.88	-8.03	0.066
(b)	TMLE	1.00	5.78	0.026	1.00	1.51	0.028	1.00	-1.46	0.075
	AIPW	1.06	9.30	0.027	1.04	4.62	0.029	1.00	1.66	0.075
	IPW	1.65	27.05	0.036	1.77	17.54	0.045	1.46	11.51	0.106
	Firpo	1.63	26.12	0.036	1.74	15.83	0.045	1.45	8.16	0.107
	OD	0.83	-8.47	0.021	1.06	-9.71	0.028	0.91	-8.03	0.066
(c)	TMLE	1.00	3.09	0.029	1.00	-0.42	0.031	1.00	-1.44	0.077
	AIPW	1.06	7.15	0.030	2.22	0.59	0.068	1.83	4.94	0.141
	IPW	1.82	-2.82	0.052	2.99	-5.71	0.092	1.60	-17.35	0.117
	Firpo	1.29	-0.48	0.037	1.55	-1.14	0.048	1.42	0.14	0.110
	OD	0.99	-21.33	0.020	1.42	-27.93	0.026	1.04	-23.48	0.060
(d)	TMLE	1.00	10.06	0.028	1.00	3.90	0.029	1.00	1.19	0.074
	AIPW	1.08	14.24	0.029	1.07	7.75	0.030	1.04	5.26	0.075
	IPW	1.50	27.05	0.036	1.70	17.54	0.045	1.48	11.51	0.106
	Firpo	1.49	26.12	0.036	1.67	15.83	0.045	1.47	8.16	0.107
	OD	0.97	-21.33	0.020	1.47	-27.93	0.026	1.09	-23.48	0.060

Table 2: Simulation results for different scenarios for the initial estimators (Scen.). % Bias is the bias relative to the true parameter value. RMSE is the MSE relative to the MSE of the TMLE.

n	% Bias		$\sqrt{n} \times \text{MSE}$		Power	
	Mean	Median	Mean	Median	Mean	Median
5,000	-3.541	2.178	5.202	2.621	0.649	0.944
10,000	-4.082	0.787	4.750	2.605	0.894	0.999
20,000	-4.283	0.683	5.302	2.387	0.982	1.000
40,000	-4.566	0.490	5.843	2.355	0.995	1.000

Table 3: Simulation results comparing TMLE of the effect on the mean vs the effect on the median as a measure of the causal effect of treatment among the treated.

comparing the performance of estimators of the effect of the mean and the effect on the median as test statistics for this hypothesis. For the effect on the mean, we focus on the

TMLE for the average treatment effect on the treated presented in Chapter 8 of [van der Laan and Rose \(2011\)](#). This estimator provides a fair competitor for our estimator of the effect on the median among the treated, since it is also doubly robust and locally efficient in the non-parametric model. A comparison of the relative MSE of the two estimators provides the increase in sample size necessary to obtain comparable power.

We focus on a scenario with both models for G_0 and e_0 correctly specified. In this scenario, in light of Theorem 1, we can use the empirical variance $\widehat{\text{Var}}\{D_{\hat{\eta}, \hat{\theta}_{\text{TMLE}}}(Z)\}$ of the estimated efficient influence function as a consistent estimator of the TMLE. This, together with the asymptotic normality of $\hat{\theta}_{\text{TMLE}}$, allows us to perform Wald-type hypothesis tests of no treatment effect on the median. A result analogous to Theorem 1 is available for the TMLE of the effect on the mean among the treated (see [van der Laan and Rose, 2011](#), for details). Simulation scenarios with misspecification of either model would require estimation of the variance through the bootstrap and would imply prohibitive computation times. Table 3 contains a comparison between both estimators in terms of their percent bias, the squared root of the mean squared error scaled by n , and the power for testing the hypothesis of no treatment effect.

Note the important loss of power for the test based on the mean as compared to its median counterpart. The hypothesis test based on the mean requires at least 2 times the sample size to achieve the power obtained with the test based on the median. This is very relevant in our setting since the sample size is not subject to modification but rather fixed by the number of AdWords customers available in a certain time period. In addition, the MSE of the estimator for the mean effect scaled by n seems to be increasing. Because the estimator used is asymptotically efficient, a value of $n \times \text{MSE}$ that diverges is a strong indication that the effect on the mean is not estimable at a \sqrt{n} -rate.

6 Concluding Remarks

The TMLE algorithm proposed here is one of many possible ways to obtain a solution to estimating equation (4). Algorithms that aim at directly solving the relevant estimating equation have been considered before. One option, discussed by [Chaffee and van der Laan \(2011\)](#) is to directly estimate the parameter ϵ as the solution to the estimating equation. A second option is to optimize the log-likelihood function constrained to the set of parameters that solve (4). We favor the algorithm presented because, in addition to solving the estimating equation, it guarantees an increase in the likelihood of the final estimate \tilde{P} compared to the initial \hat{P} . This has been shown to yield estimators with improved finite sample properties ([Chaffee and van der Laan, 2011](#)).

When the dimension of the baseline variables is large relative to the sample size, the curse of dimensionality precludes the use of nonparametric estimators for e_0 and G_0 ([Robins and Ritov, 1997](#)). A potential way to address this is to incorporate data-adaptive model selection in constructing the initial estimators in step 1 of the TMLE procedure, such as model stacking ([Wolpert, 1992](#)) or super learning ([van der Laan et al., 2007](#)). The asymptotic normality of our estimator then require conditions (ii)-(iii) in Theorem 1. These conditions would hold automatically for the MLE in a parametric model, but need to be verified for data-adaptive estimators. [van der Laan \(2014\)](#) proposed an estimator for the case of a the mean in a missing data model that relaxes assumption (ii); this approach is generalizable to our problem.

Appendix A Proofs

A.1 Lemma 2

Proof. By the law of iterated expectation we have

$$\begin{aligned} P_0 D_{\eta, \theta} &= -\frac{1}{f(\theta)} P_0 \left[\frac{e_0}{e} (h_{0, \theta} - h_\theta) + h_\theta - q \right] \\ &= -\frac{1}{f(\theta)} P_0 \left[\left(\frac{e_0}{e} - 1 \right) (h_{0, \theta} - h_\theta) + (h_{0, \theta} - q) \right], \end{aligned} \quad (7)$$

where $h_{0, \theta}(x) = G_0(\theta \mid 1, x)$ and $h_\theta(x) = G(\theta \mid 1, x)$. The lemma follows from substituting either $G = G_0$ or $e = e_0$ in the previous display, with $\theta = \theta_0$. \square

A.2 Theorem 1

Proof. By construction of the TMLE algorithm we have $\mathbb{P}_n D_{\tilde{\eta}, \hat{\theta}_{\text{TMLE}}} = o_P(1/\sqrt{n})$. By Lemma 2 and Assumption (i) we have $P_0 D_{\eta_1, \theta_0} = 0$. In addition, by Theorem 5.9 of [van der Vaart \(2000\)](#) we have $\hat{\theta}_{\text{TMLE}} = \theta_0 + o_P(1)$. Under assumptions (i)-(iii), an application of Theorem 5.31 of [van der Vaart \(2000\)](#) yields

$$\sqrt{n}(\hat{\theta}_{\text{TMLE}} - \theta_0) = \sqrt{n} P_0 D_{\tilde{\eta}, \theta_0} + \sqrt{n}(\mathbb{P}_n - P_0) D_{\eta_1, \theta_0} + o_P(1 + \sqrt{n} |P_0 D_{\tilde{\eta}, \theta_0}|). \quad (8)$$

Let the influence function of Assumption (ii) be denoted by Δ . Then we have $\sqrt{n} P_0 D_{\tilde{\eta}, \theta_0} = \sqrt{n}(\mathbb{P}_n - P_0) \Delta + o_P(1)$. Thus

$$\begin{aligned} \sqrt{n}(\hat{\theta}_{\text{TMLE}} - \theta_0) &= \sqrt{n}(\mathbb{P}_n - P_0)(\Delta + D_{\eta_1, \theta_0}) + o_P(1 + O_P(1)) \\ &= \sqrt{n}(\mathbb{P}_n - P_0)(\Delta + D_{\eta_1, \theta_0}) + o_P(1). \end{aligned}$$

The central limit theorem yields the claimed asymptotic normality.

Efficiency under the assumption that $\eta_1 = \eta_0$ follows from the following argument. Equations (7) and the Cauchy-Schwartz inequality yields

$$P_0 D_{\tilde{\eta}, \theta_0} \leq C \|\tilde{h}_{\theta_0} - h_{0, \theta_0}\| \|\hat{e} - e_0\|,$$

for some constant C . Under assumption (i) the term in the right hand side is $o_P(1/\sqrt{n})$. Then, from Equation (8) it follows that

$$\sqrt{n}(\hat{\theta}_{\text{TMLE}} - \theta_0) = \sqrt{n}(\mathbb{P}_n - P_0)D_{\eta_0, \theta_0} + o_P(1),$$

so that $\hat{\theta}_{\text{TMLE}}$ is asymptotically normal, consistent, and efficient. \square

Appendix B Super Learning for a Conditional Density

The conditional density g_0 may be defined as the minimizer of the negative log-likelihood loss function. That is $g_0 = \arg \min_{f \in \mathcal{F}} R(f, p_0)$, where \mathcal{F} is the space of all non-negative functions of (y, x) satisfying $\int f(y, x) dy = 1$, and $R(f) = -\int m \log f(y, x) dP_0(z)$. An estimator \hat{g} is seen here as an algorithm that takes a training sample $\mathcal{T} \subseteq \{Z_i : i = 1, \dots, n\}$ as an input, and outputs an estimated function $\hat{g}(y \mid 1, x)$.

For a given estimator \hat{g} , we use cross-validation to construct an estimate $\hat{R}(\hat{g})$ of the risk $R(\hat{g})$ as follows. Let $\mathcal{V}_1, \dots, \mathcal{V}_J$ denote a random partition of the index set $\{1, \dots, n\}$ into J validation sets of approximately the same size. That is, $\mathcal{V}_j \subset \{1, \dots, n\}$; $\bigcup_{j=1}^J \mathcal{V}_j = \{1, \dots, n\}$; and $\mathcal{V}_j \cap \mathcal{V}_{j'} = \emptyset$. In addition, for each j , the associated training sample is given by $\mathcal{T}_j = \{1, \dots, n\} \setminus \mathcal{V}_j$. Denote by $\hat{g}_{\mathcal{T}_j}$ the estimated density function obtained by training the algorithm using only data in the sample \mathcal{T}_j . The cross-validated risk of an estimated

density \hat{g} is defined as

$$\hat{R}(\hat{g}) = -\frac{1}{J} \sum_{j=1}^J \frac{1}{|\mathcal{V}_j|} \sum_{i \in \mathcal{V}_j} \log M_i \hat{g}_{\mathcal{T}_j}(Y_i | 1, X_i). \quad (9)$$

Consider now a finite collection $\mathcal{L} = \{\hat{g}_k : k = 1, \dots, K_n\}$ of candidate estimators for g_0 . We call this collection a *library*. We define the stacked predictor as a convex combination of the predictors in the library:

$$\hat{g}_\alpha(y | 1, x) = \sum_{k=1}^{K_n} \alpha_k \hat{g}_k(y | 1, x),$$

and estimate the weights α as the minimizer of the cross-validated risk $\hat{\alpha} = \arg \min \hat{R}(\hat{g}_\alpha)$, subject to $\sum_{k=1}^{K_n} \alpha_k = 1$. The final estimator is then defined as $\hat{g}_{\hat{\alpha}}$.

Construction of the library Consider a partition of the range of A into k bins defined by a sequence of values $\beta_0 < \dots < \beta_k$. Consider a candidate for estimation of $g_0(y | 1, w)$ given by

$$\hat{g}_\beta(y | 1, x) = \frac{\widehat{Pr}(Y \in [\beta_{t-1}, \beta_t) | M = 1, X = x)}{\beta_t - \beta_{t-1}}, \text{ for } \beta_{t-1} \leq y < \beta_t. \quad (10)$$

Here \widehat{Pr} denotes an estimator of the true probability $Pr_0(Y \in [\beta_{t-1}, \beta_t) | M = 1, X = x)$ obtained through a hazard specification and the use of an estimator for the expectation of a binary variable in a repeated measures dataset as follows. Consider the following

factorization

$$Pr(Y \in [\beta_{t-1}, \beta_t] | M = 1, X = x) = Pr(Y \in [\beta_{t-1}, \beta_t] | Y \geq \beta_{t-1}, M = 1, X = x) \times \prod_{j=1}^{t-1} \{1 - Pr(Y \in [\beta_{j-1}, \beta_j] | Y \geq \beta_{j-1}, M = 1, X = x)\}.$$

The likelihood for model (10) is proportional to

$$\prod_{i=1}^n Pr(Y_i \in [\beta_{t-1}, \beta_t] | M = 1, X) = \prod_{i=1}^n \left[\prod_{j=1}^{t-1} \{1 - Pr(Y_i \in [\beta_{j-1}, \beta_j] | Y_i \geq \beta_{j-1}, M_i = 1, X_i)\} \right] \times Pr(Y_i \in [\beta_{t-1}, \beta_t] | Y_i \geq \beta_{t-1}, M_i = 1, X_i),$$

which corresponds to the likelihood for the expectation of the binary variable $I(Y_i \in [\beta_{j-1}, \beta_j])$ in a repeated measures data set in which the observation of subject i is repeated k times, conditional on the event $Y_i \geq \beta_{j-1}$.

Thus, each candidate estimator for g_0 is indexed by two choices: the sequence of values $\beta_0 < \dots < \beta_k$, and the algorithm for estimating the probabilities $Pr_0(Y_i \in [\beta_{t-1}, \beta_t] | Y_i \geq \beta_{t-1}, M_i = 1, W_i)$. The latter is simply a conditional probability, and therefore any standard prediction algorithm may be used as a candidate. In the remainder of this section we focus on the selection of the location and number of bins, implied by the choice of β_j values.

Denby and Mallows (2009) describe the histogram as a graphical descriptive tool in which the location of the bins can be characterized by considering a set of parallel lines cutting the graph of the empirical cumulative distribution function (ECDF). Specifically, given a number of bins k , the equal-area histogram can be regarded as a tool in which the ECDF graph is cut by $k + 1$ equally spaced lines parallel to the x axis. The usual

equal-bin-width histogram corresponds to drawing the same lines parallel to the y axis. In both cases, the location of the cutoff points for the bins is defined by the x values of the points in which the lines cut the ECDF. As pointed out by the authors, the equal-area histogram is able to discover spikes in the density, but it oversmooths in the tails and is not able to show individual outliers. On the other hand, the equal-bin-width histogram oversmooths in regions of high density and does not respond well to spikes in the data, but is a very useful tool for identifying outliers and describing the tails of the density.

As an alternative to find a compromise between these two approaches, the authors propose a new histogram in which the ECDF is cut by lines $x + cy = bh$, $b = 1, \dots, k + 1$; where c and h are parameters defining the slope and the distance between lines, respectively. The parameter h identifies the number of bins k . The authors note that $c = 0$ gives the usual histogram, whereas $c \rightarrow \infty$ corresponds to the equal-area histogram.

Thus, we can define a library of candidate estimators for the conditional density in terms of (10) by defining values of the vector β through different choices of c and k , and considering a library for estimation of conditional probabilities. Specifically, the library is given by the Cartesian product

$$\mathcal{L} = \{c_1, \dots, c_{m_c}\} \times \{k_1, \dots, k_{m_k}\} \times \{\widehat{Pr}_1, \dots, \widehat{Pr}_{m_P}\},$$

where the first is a set of m_c candidate values for c , the second is a set of m_k candidate values for k , and the third is a set of m_P candidates for the probability estimation algorithm. The use of this approach will result in estimators that are able to identify regions of high density as well as provide a good description of the tails and outliers of the density. The inclusion of various probability estimators allows the algorithm to find possible nonlinearities and higher order interactions in the data. This proposed library may be augmented by considering any other estimator. For example, there may be expert knowledge leading to believe that a

normal distribution (or any other distribution) with linear conditional expectation could fit the data. A candidate algorithm that estimates such a density using maximum likelihood may be added to the library. This algorithm was first proposed by [Díaz and van der Laan \(2011\)](#), the reader interested in more details and applications is encouraged to consult the original research article.

Appendix C R Code

```
trim <- function(x) pmax(x, 1e-10)

compute.quantile <- function(Q, w, q, r){
  F <- function(y)sapply(y, function(x)mean(rowSums((Q <= x) * w)))
  inv <- function(qq){
    uniroot(function(x){F(x) - qq}, r, extendInt = 'yes')$root
  }
  return(sapply(q, function(qq)inv(qq)))
}

od <- function(y, t, Q, g, q){
  n <- length(y)
  w <- matrix(1/dim(Q)[2], ncol = dim(Q)[2], nrow = n)
  chiq <- compute.quantile(Q, w, q, range(y))
  return(chiq)
}

tmle <- function(y, t, Q, g, q){
  n <- length(y)
  D <- function(y, w, chiq){
    1 / g * ((y <= chiq) - rowSums((Q <= chiq) * w))
  }
  w <- matrix(1/dim(Q)[2], ncol = dim(Q)[2], nrow = n)
  h <- t
  chiq <- compute.quantile(Q, w, q, range(y))
  Do <- D(y, w, chiq)
  Dq <- D(Q, w, chiq)
  iter <- 1
  crit <- TRUE
  max.iter <- 20
  while(crit && iter <= max.iter){
    est.eq <- function(eps){
```

```

    out <- - mean(h * (Do - rowSums(Dq * exp(eps * Dq) * w) /
    rowSums(exp(eps * Dq) * w)))
    return(out)
  }
  loglik <- function(eps){
    out <- - mean(h * (eps * Do - log(rowSums(exp(eps * Dq) * w))))
    return(out)
  }
  eps <- optim(par = 0, loglik, gr = est.eq, method = 'BFGS')$par
  w <- exp(eps * Dq) * w / rowSums(exp(eps * Dq) * w)
  chiq <- compute.quantile(Q, w, q, range(y))
  Do <- D(y, w, chiq)
  Dq <- D(Q, w, chiq)
  iter <- iter + 1
  crit <- abs(eps) > 1e-4 / n^0.6
}
return(chiq)
}

firpo <- function(y, t, Q, g, q){
  library(quantreg)
  h <- t / g
  chiq <- coef(rq(y ~ 1, weights = h, tau = q))
  names(chiq) <- NULL
  return(chiq)
}

ipw <- function(y, t, Q, g, q){
  n <- length(y)
  w <- matrix(1/dim(Q)[2], ncol = dim(Q)[2], nrow = n)
  h <- t / g
  D <- Vectorize(function(chiq) mean(h * (y <= chiq) - q))
  chiq <- uniroot(D, c(-1000, 1000), extendInt = 'yes')$root
  return(chiq)
}

aipw <- function(y, t, Q, g, q){
  n <- length(y)
  w <- matrix(1/dim(Q)[2], ncol = dim(Q)[2], nrow = n)
  h <- t / g
  D <- Vectorize(function(chiq){
    mean(h * ((y <= chiq) - rowSums((Q <= chiq) * w))) +
    mean(rowSums((Q <= chiq) * w) - q)
  })
  chiq <- uniroot(D, c(-1000, 1000), extendInt = 'yes')$root
  return(chiq)
}

```



```

}

datagen <- function(n) {
  kBeta <- c(210, 27.4, 13.7, 13.7, 13.7)
  kTheta <- c(-1, 0.5, -0.25, -0.1)
  Z <- matrix(rnorm(n * 4), nrow = n, ncol = 4)
  X <- matrix(nrow = n, ncol = 4)
  X[, 1] <- exp(Z[, 1] / 2)
  X[, 2] <- Z[, 2] / (1 + exp(Z[, 1])) + 10
  X[, 3] <- (Z[, 1] * Z[, 3] / 25 + 0.6) ^ 3
  X[, 4] <- (Z[, 2] + Z[, 4] + 20) ^ 2
  y <- rnorm(n, mean = cbind(1, Z) %*% kBeta)
  true.prop <- 1 / (1 + exp(-Z %*% kTheta))
  T <- rbinom(n, 1, true.prop)
  dat <- list(Y = y, T = T, Z = Z, X = X)
  return(dat)
}

## Kang & Schafer Example

n.quant <- 500
formT <- T ~ X1+X2+X3+X4
formY <- Y ~ X1+X2+X3+X4
data <- datagen(1000)
X <- data$X
Y <- data$Y
T <- data$T
fitT <- glm(formT, data = data.frame(T=T, X), family = binomial)
fitY1 <- lm(formY, data = data.frame(Y=Y, T=T, X), subset = T == 1)
fitY0 <- lm(formY, data = data.frame(Y=Y, T=T, X), subset = T == 0)
median1 <- predict(fitY1, newdata = data.frame(T=1, X))
median0 <- predict(fitY0, newdata = data.frame(T=0, X))
Q1 <- sapply(seq(1/n.quant, 1 - 1/n.quant, 1/n.quant),
             function(q) qnorm(q, mean = median1, sd = summary(fitY1)$sigma))
Q0 <- sapply(seq(1/n.quant, 1 - 1/n.quant, 1/n.quant),
             function(q) qnorm(q, mean = median0, sd = summary(fitY0)$sigma))
g1 <- trim(predict(fitT, type = 'response'))
ame <- tmle(Y, T, Q1, g1, q) - tmle(Y, 1 - T, Q0, 1 - g1, q)

```

References

- Peter J. Bickel, Chris A.J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, 1997.
- Moshe Buchinsky. Recent advances in quantile regression models: a practical guideline for empirical research. *Journal of human resources*, pages 88–126, 1998.
- Paul Chaffee and Mark J van der Laan. Targeted minimum loss based estimation based on directly solving the efficient influence curve equation. 2011.
- P.E. Cheng and C.K. Chu. Kernel estimation of distribution functions and quantiles with missing data. *Statistica Sinica*, 6:63–78, 1996.
- Victor Chernozhukov, Iván Fernández-Val, and Blaise Melly. Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268, 2013.
- L. Denby and C. Mallows. Variations on the histogram. *Journal of Computational and Graphical Statistics*, Vol. 18, Iss. 1:21–31, 2009.
- Iván Díaz and Mark van der Laan. Super learner based conditional density estimation with application to marginal structural models. *The International Journal of Biostatistics*, 7(1):38, 2011.
- Sergio Firpo. Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, pages 259–276, 2007.
- Markus Frölich and Blaise Melly. Unconditional quantile treatment effects under endogeneity. *Journal of Business & Economic Statistics*, 31(3):346–357, 2013.
- Susan Gruber and Mark J van der Laan. A targeted maximum likelihood estimator of a

- causal effect on a bounded continuous outcome. *The International Journal of Biostatistics*, 6(1), 2010.
- Zonghui Hu, Dean A Follmann, and Jing Qin. Dimension reduced kernel estimation for distribution function with incomplete data. *Journal of statistical planning and inference*, 141(9):3084–3093, 2011.
- Joseph DY Kang and Joseph L Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, pages 523–539, 2007.
- Roger Koenker. *Quantile regression*. Number 38. Cambridge university press, 2005.
- Roger Koenker. *quantreg: Quantile Regression*, 2013. URL <http://CRAN.R-project.org/package=quantreg>. R package version 5.05.
- Xu Liu, Peixin Liu, and Yong Zhou. Distribution estimation with auxiliary information for missing data. *Journal of Statistical Planning and Inference*, 141(2):711–724, 2011.
- Blaise Melly. Estimation of counterfactual distributions using quantile regression. *Review of Labor Economics*, 68(4):543–572, 2006.
- Whitney K. Newey. Semiparametric efficiency bounds. *Journal of applied econometrics*, 5(2):99–135, 1990.
- Eric C. Polley, Sherri Rose, and Mark J. van der Laan. Super learning. In Mark J. van der Laan and Sherri Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Data*, pages 43–66. Springer New York, New York, NY, 2011.
- Kristin E Porter, Susan Gruber, Mark J van Der Laan, and Jasjeet S Sekhon. The relative

- performance of targeted maximum likelihood estimators. *The International Journal of Biostatistics*, 7(1):1–34, 2011.
- James Robins, Mariela Sued, Quanhong Lei-Gomez, and Andrea Rotnitzky. Comment: Performance of double-robust estimators when inverse probability weights are highly variable. *Statist. Sci.*, 22(4):544–559, 11 2007a.
- James Robins, Mariela Sued, Quanhong Lei-Gomez, and Andrea Rotnitzky. Comment: Performance of double-robust estimators when” inverse probability” weights are highly variable. *Statistical Science*, 22(4):544–559, 2007b.
- James M Robins and Ya’acov Ritov. Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16(3):285–319, 1997.
- Richard JCM Starman. Models, inference, and truth: probabilistic reasoning in the information era. In Mark van der Laan and Sherri Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, 2011.
- Ori M Stitelman, Victor De Gruttola, and Mark J van der Laan. A general implementation of tmle for longitudinal data applied to causal inference in survival analysis. *The international journal of biostatistics*, 8(1), 2012.
- Mark J van der Laan. Targeted estimation of nuisance parameters to obtain valid statistical inference. *The International Journal of Biostatistics*, 10(1):29–57, 2014.
- Mark J. van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.
- Mark J. van der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

- Mark J. van der Laan, Eric Polley, and Alan Hubbard. Super learner. *Statistical Applications in Genetics & Molecular Biology*, 6(25), 2007. ISSN 1.
- Aad W. van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.
- Richard von Mises. On the asymptotic distribution of differentiable statistical functions. *The annals of mathematical statistics*, pages 309–348, 1947.
- Qihua Wang and Yongsong Qin. Empirical likelihood confidence bands for distribution functions with missing responses. *Journal of Statistical Planning and Inference*, 140(9): 2778–2789, 2010.
- David H Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.
- Keming Yu and MC Jones. Local linear quantile regression. *Journal of the American statistical Association*, 93(441):228–237, 1998.
- Zhiwei Zhang, Zhen Chen, James F Troendle, and Jun Zhang. Causal inference on quantiles with an obstetric application. *Biometrics*, 68(3):697–706, 2012.
- Pu-Ying Zhao, Man-Lai Tang, and Nian-Sheng Tang. Robust estimation of distribution functions and quantiles with non-ignorable missing data. *Canadian Journal of Statistics*, 41(4):575–595, 2013.